Artificial Intelligence: CS F407 Assignment

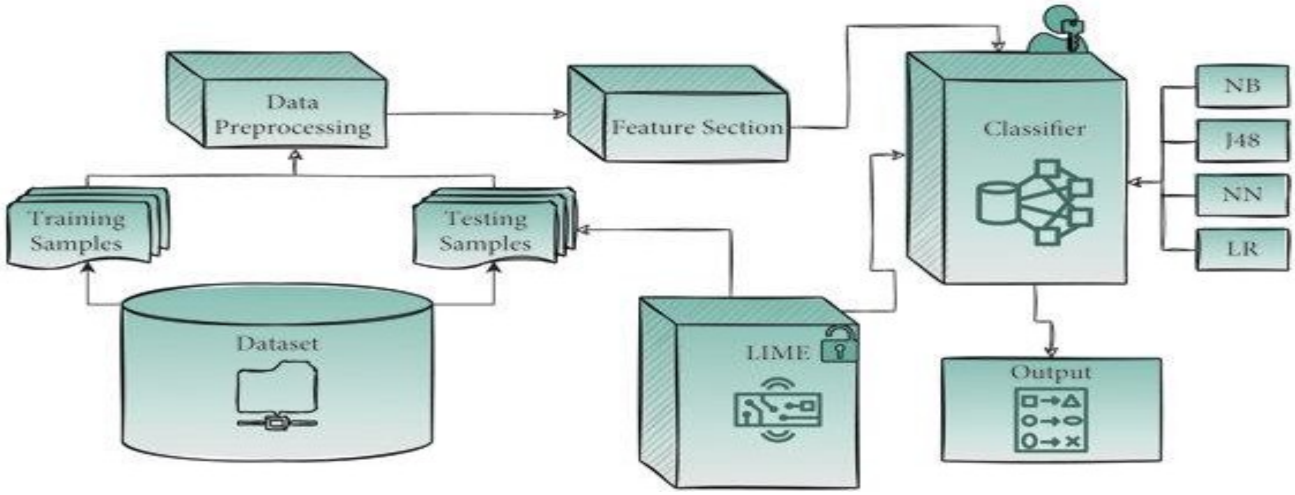
BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

ज्ञानं परमं बलम्

NAME: CHIRANJEEV SINGH

ID: 2022A1PS1631P

Domain: *Placement Prediction System for students*



I. Abstract

The evolving field of AI offers vast opportunities in various fields yet it poses significant challenges in predicting system (placement prediction system for students or job prediction for user with specific skill set or model for STEM/ Non- STEM Job prediction).

This assignment explores the placement prediction system for students aimed at enhancing career counseling, job selections and educational planning using various research methods. On analyzing different papers we come across tools like data mining using classification algorithms (WEKA)-Naïve Bayse, C4,5 Tree, Multilayer Perception or using DKT(deep knowledge tracing and enhanced DKT or ML model constructions for placement prediction using algorithms like Logistic Regression, SVM, XG boost with accuracy of 88%, 91% and 84% respectively for each algorithm. However, each such method has its own limitations such as quality of data set for data mining, overfitting and dimensionality issues.

The aim of the assignment is to interpret a suitable prediction system model for student's placements based on different attributes like user skillset, click stream record, student profile, specialization in courses, unique skills, educational background and so on with high efficiency and low errors than the other models resulting in quick, accurate and real time decision making. This model can determine the relations between academic achievement of students and their placement in campus selection

The assignment requires a thorough study of the research papers in the selected domain. *The key dataset is a user-skill matrix of 215 students records with 15 features with target label with frequency counts for the attributes they possess like gender, percentage of high school, stream, work experience, stream in degree, specialization etc*. Preprocessing sanitizes and normalizes the raw skill data extracted from user profiles. This provides structured input data for the collaborative filtering approach to generate job recommendations.

Addressing these limitations and gaps not only improve the accuracy of predictions but also supports students in making well-informed career decisions ultimately contributing to better alignment between educational output and economical needs. This lays the groundwork for future research aimed at developing improved predictive system that is both comprehensive, reliable and adaptive to individual and economic shifts.

Keywords: Data mining, DKT, WEKA, LR, SVM, Machine learning

II. Introduction

Education system is backbone of progressing Indian society. India’s superiority in information technology is due to evolution in higher education system in last few years. One has to concentrate on challenges faced by higher educational institutes to improve quality of higher education. Majority of students in higher education join a course for securing a good job. Placements play a crucial role in every educational institution. College reputation is based on students’ placement rate. The main goal of institutes is to get their students placed with a better offer. Every educational institution is mainly focused on student placements. Educational institutions are fulfilling student dreams using placements. A placement prediction system can be used to determine the capability of a student for that specific job role

Manual prediction needs lots of human resources to maintain student records With the machine learning algorithms, we overcome the problem in the manual process. Placement prediction system helps in effective filtering of students by considering various factors like CGPA, technical skills, soft skills, coding skills, communication skills. Using different Machine Learning (ML) algorithms, predict the probability of student placement performance. In this we used different machine learning algorithms like support vector machine (SVM), Logistic Regression (LR), Naive Bayes, XG Boost, Decision Tree to predict the student performance. These classifiers independently predict the results and then compare the accuracy of the algorithms, on the data set. A Machine Learning system trains from past data, constructs the forecast models, and on getting new data, predicts the output for it. The accuracy of expected output is dependent on the amount of data, large amounts of data aid in more precise forecasting.

The assignment is organized as follows: Section I presents an abstract- a brief overview, Section II is introduction which describes background details and approach to the assignment problem, Section III describes the literature review, Section IV describes the assignment problem, Section V and VI focuses on implementation of the problem and presentation and analysis of all collected results. Finally we conclude this assignment with suggestion for future work and references in Section VII and VIII.

III. Literature Review:

This section reviews various predictive models employed to forecast student placements. The study focuses on identifying common methodologies used, evaluating the effectiveness, significance and limitations.

The research paper used for the assignment along with its summary , algorithms, limitations and metrices are listed as follows:

Model Construction Using ML for Prediction of Student Placement:
International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 10 Issue VI June 2022

Machine learning algorithms estimate new output values using historical data as input. A Machine Learning system trains from past data, constructs the forecast models, and on getting new data, predicts the output for it. The accuracy of expected output is dependent on the amount of data, large amounts of data aid in more precise forecasting. If we have a complex situation for which we need to make predictions, rather than writing code for it, we may just input the data to generic algorithms, and the machine will develop the logic based on the data and forecast the outcome. Machine learning has shifted our perspective on the issue.

Types of ML classification:

1) Supervised Machine Learning
2) Unsupervised learning
3) Reinforcement

Methodology:

- Logistic Regression Logistic regression algorithm is used to predict the dependent variable on a given set of independent variables. log(a/1-a) is the link function. It predicts the dependent variable in a categorical form (in the form of binary 0 or 1, or yes/no).
- Naive Bayes Classier - Gaussian Naive Bayes Bayes' Theorem also called as Bayes' law or Bayes' rule. It says the probability of the occurrence of an event based on the knowledge it has on that event.
- XG Boost XGBoost stands for Extreme Gradient Boosting. It is a gradient boosted ML library used to find accurate model based on the data.
- Decision Tree Decision Tree is a widely used classification as well as Regression problem. It is a Supervised learning technique. In decision

tree internal nodes represent the features of a dataset, branches depict the decision rules and each leaf node says the outcome.

- Information Gain Information gain nothing but the reduction of uncertainty and it also says which attribute should be selected

Life cycle of ML:

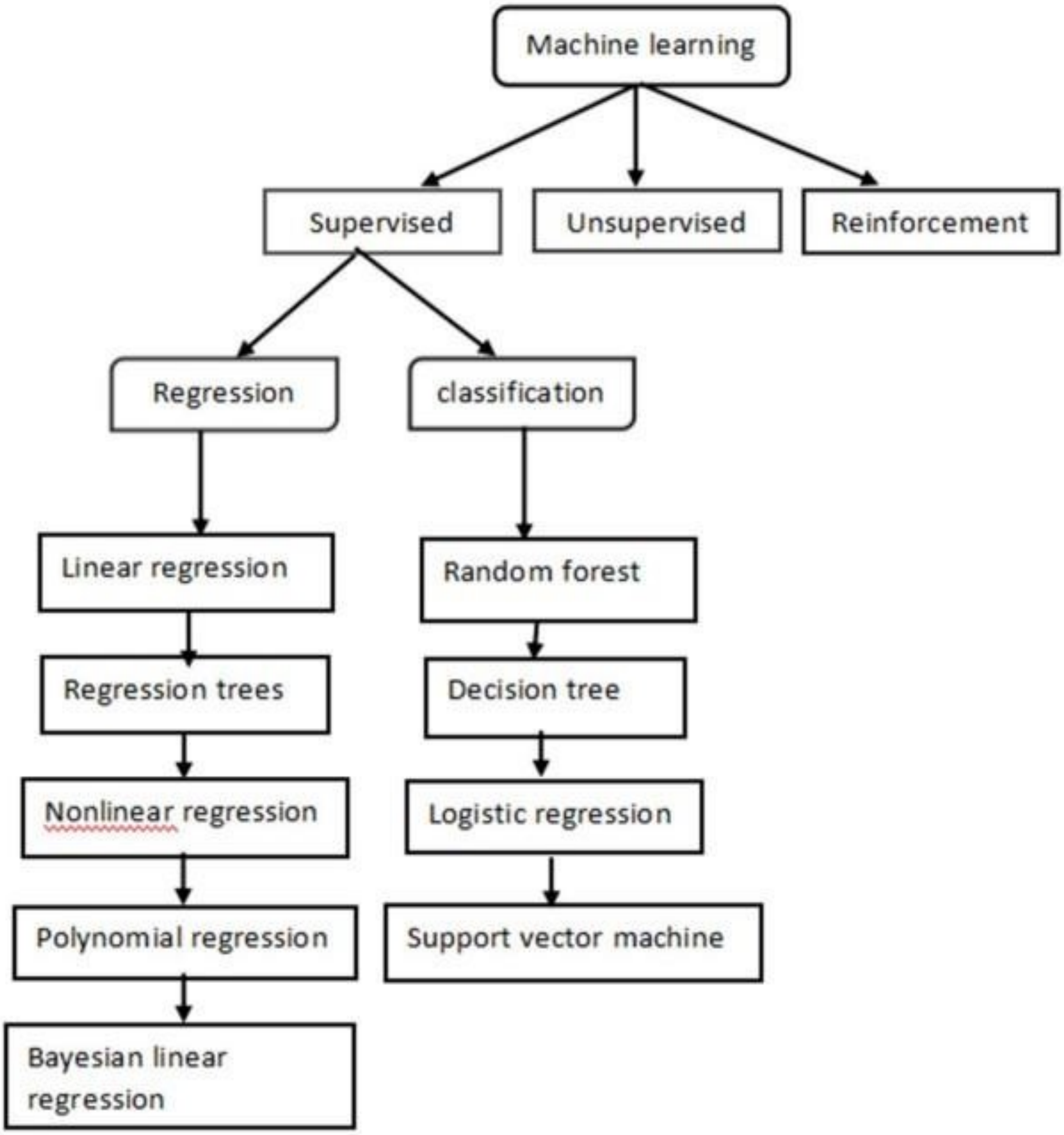


Fig 1: Machine Learning Taxanomy

Limitations:

1) Assumption of Linearity: LR assumes a linear relationship between input features and the log-odds of the outcome. This assumption may not hold true for complex, non-linear relationships.
2) Limited Expressiveness: Naïve Bayes assumes feature independence, this assumption can lead to biased estimations.
3) Sensitive to Imbalanced data: All the ML models can be sensitive to imbalanced datasers where one is significantly more prevalent thsn the others.

Some other limitations are Difficulty Handling Categorical features, Overfitting, Scalability issue, limited representation of Uncertainty

IV. Assignment Problem

Developing a Placement System for students using ML models and analyse the results from the dataset to determine the accuracy and efficiency of the model

Task:

1)Data collection and Processing

2)Model Selection and development

3)Model Evaluation and Validation

4)Hyperparameter Tuning and Optimization

5)Deployment and analysis

V)Implementation

Implementing a placement prediction system for students involves several steps from data preprocessing to model training and deployment.

Various implementation of the ML algorithms are follows (the code for the models is attached with the file)

- **Logistic Regression:** Logistic Regression Logistic regression algorithm is used to predict the dependent variable on a given set of independent variables. log(a/1-a) is the link function. It predicts the dependent variable in a categorical form (in the form of binary 0 or 1, or yes/no).

  $$\log [a/1-a]= Y$$

  The generalized linear model equation is:

  $$G1(E1(y1) = \alpha 1+ \beta x1+ \gamma x2$$

  Here, g1() is the link function, E1(y1) is the target variable's expectation and $\alpha 1+ \beta x1 + \gamma x2$ is the linear predictor ($\alpha 1$, β, γ to be predicted) is the linear predictor ($\alpha 1$, β, γ to be predicted).
- **Naive Bayes Classifier** - It says the probability of the occurrence of an event based on the knowledge it has on that event. Mathematical equation for Bayes' Theorem is:

  $$P(S/T) =P(T/S).\ P(S)/P(T)$$

  where S and T are events and $P(T) \neq 0$. P(S|T) is called as the Posterior probability which means Probability of hypothesis S on the observed event T. P(T|S) is called as the Likelihood probability which means Probability of the evidence given that the probability of a hypothesis is true. P(S) is called as the Prior Probability which means Probability of hypothesis before observing the evidence. P(T) is called as the Marginal Probability which means Probability of Evidence
- **XG Boost-** XG Boost stands for Extreme Gradient Boosting. It is a gradient boosted ML library used to find accurate model based on the data.

  $$G2(X)= \text{sigma}(0+1+ f1(X) + 1 + f2(X))$$

  where G2(x) value is taken as the prediction from Boost model. G0, the initial model is defined to predict the target variable y. A new model f1 is fit to the residuals from the previous step. Now, G0 and f1 are combined to give G1, the boosted version of G0. The mean squared error from G1 will be lower than that from G0. Now, the deep workings of XG Boost is given below.

  $$G1(X) < -G0(X) +f1(X)\_$$

To improve G1 performance, we could model after the residuals of G1 and create a new model

G2(X) < -G1(X) + f2(X)

Let this be done for 'n' iterations, until residuals have been decreased as much as possible:

Gn(X) < -Gn -1(X) +fm(X)

G1(X) < -G0(X) + f1(X)

- **Decision Tree** -Decision Tree is a widely used classification as well as Regression problem. It is a Supervised learning technique. In decision tree internal nodes represent the features of a dataset, branches depict the decision rules and each leaf node says the outcome.
  Entropy = En(S1)= -p1(+) logp1(+) – p1(-)log p1(-)
  where p1+ is the probability of positive class, p1– is the probability of negative class

- **Information Gain**- Information gain nothing but the reduction of uncertainty and it also says which attribute should be selected.
  Info Gain = En(Y1)- En(Y1|X1|)

- **Random Forest-** Random Forest belongs to the ensemble learning family, which combines predictions from multiple individual models to improve accuracy and robustness. The base model of Random Forest is the decision tree, a simple yet powerful algorithm for classification and regression tasks. Decision trees partition the feature space into regions and make predictions based on majority voting or averaging in each region. Key parameters of the Random Forest model include the number of trees (**n_estimators),** the maximum depth of the trees **(max_depth),** and the number of features to consider for splitting **(max_features**).Tuning these parameters can impact the model's performance and generalization ability

Data Set

Data on total contains 215 student records with 15 features with target label. The values of the following attributes for 215 are listed in the attached file:

Attribute	Description	Data type
sl. no	Serial number	Integer
gender	Gender of students	String
ssc_p	Percentage in intermediate	Double
ssc_b	Board of intermediate	String
hsc_p	Percentage of high school	Double
hsc_b	Board of high school	String
hsc_s	Stream in high school	String
degree_p	Percentage in degree	Double
degree_t	Stream in degree	String
workex	Work experience	String
etest_p	Percentage in e_test	Double
specialization	Specialization in mba	String
mba_p	Percentage in mba	Double
status	Status of student	String
salary	Salary offered	Integer

Data Preprocessing

Data pre-processing is a very important step, which works on the meaningless data and converts into clean data, and then trains the Machine Learning algorithms. The below mentioned are the basic steps involved in the Data Pre-Processing.

a) ***Import all the Required Libraries:*** For the language we used, if it need to identify any function we worked, we need to import all the required libraries into it. All the necessary libraries need to be imported into the working environment like Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn.

b) ***Data set Importing:*** The collected dataset which is in CSV format, need to be imported into the workspace. As the columns of sl. No and salary are not required to predict student placement status, those two columns are dropped

c) ***Dealing with the Missing Values:*** Missing values are those whose data is not present in the respective column. Dataset with more number of missing values may lead to less computational power. The dataset need to be checked for any missing values, if found then it need to be replaced by considering mean and mode of the attribute or may be either by deleting entire row or column if more number of missing values are found. In the considered dataset, found no missing values and hence not required dealing with missing values in preprocessing steps.

d) ***Label Encoding the Data:*** Label encoding is mainly done to make sure that all the data is in numerical format. The categories in the 'hsc_s' and 'degree_t' are splitted into individual columns on applying the dummy encoding. It uses '1'indicating 'YES' and '0' indicating 'NO'. Here, the number of newly created columns equals to the number of categories. The following columns ['gender','ssc_b', 'hsc_b','workex','specialisation','status'] are also converted to '0' and '1's. Outlier is an object which completely differs from the rest of the objects in the data. They cause problem for the statistical result. Outliers are checked for the data and some outliers from the columns 'hsc_p' and 'degree_p' are removed to make it more fit for the statistical analysis. Correlation is mainly checked to know the relationship between the variables. '+1' indicates that if one variable is increasing simultaneously the other is also increasing. '-1'indicates that if one variable is increasing the other is decreasing. '0' indicates that correlation is not present between the variables. Two variables x and y are considered which indicates independent and dependent variables respectively.

e) ***Dataset Splitting:*** Data set need to be divided into two sections, Training and Testing data. Training data is the major part of the dataset which is used to feed the Machine learning model to recognise the patterns. Testing data is data which is used to compute the accurate result of the model. Data can be splitted in the ratio of 70:30, 60:40, 80:20. The dataset is divided into 80:20 ratio of Training and testing respectively.

f)***Scaling the Features:*** It is the last step in any preprocessing. It is used on independent variable to limit their range using the fit transform function, so that comparison becomes easy. Standardization method is used on the dataset to limit their features. Formula for standardization:

x' = (x- mean(x))/sd

Where, x` indicates new value got, x indicates actual value, sd indicates standard deviation.

VI)Results and Analysis:

The overall analysis made us to differentiate which factors helped directly and indirectly in predicting the placement of the student and also the comparative study on the various Machine learning algorithms made us to know the most accurate algorithm which works on the student placement data. To result out the performance of classification we are required to consider the parameters such as Precision(P), Recall (R), Accuracy (Acc) and F1-score (F1_S) which confirms whether the classification is good or bad on the dataset. The classification metrics purely depends on parameters of True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN) of confusion matrix. The performance metrics are calculated as follows: Let X=True Positive, Y=True Negative, S=False Positive, T=False Negative

$$\text{Recall(Rec)} = X/X+T$$

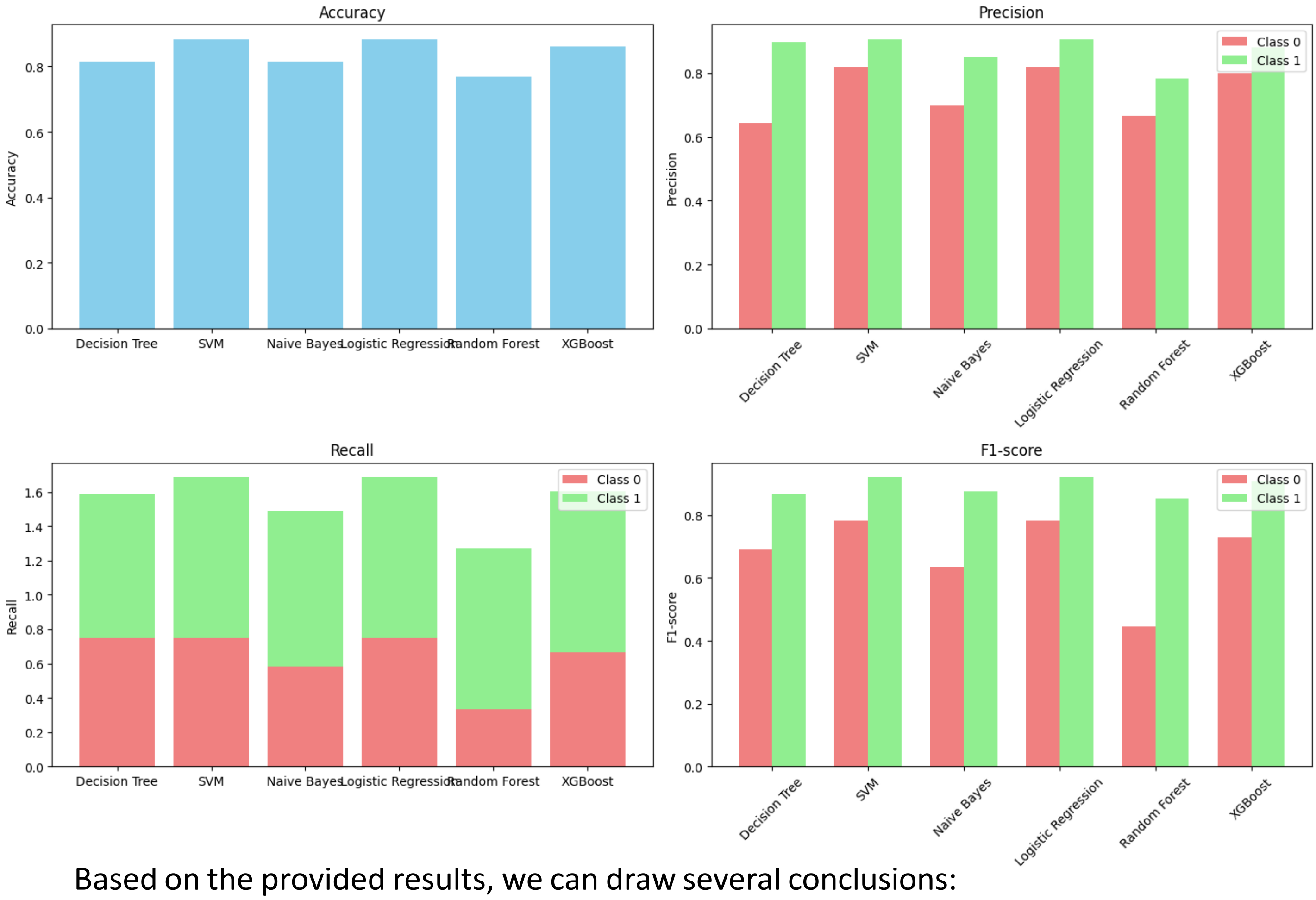
$$\text{Precision(Pre)} = (X)/(X+S)$$

$$\text{F1\_ Score} = 2.\text{Pre. rec}/\ \text{Pre} + \text{rec}$$

Here is the table with the measured data from the training set using the implementation approach:

Classifier	Accuracy	Precision(class 0)	Precision(class 1)	Recall(class0)	Recall(class1)	F1_score(class 0)	F1_score(class 0)
Decision Tree	0.8139	0.642	0.89655	0.75	0.838	0.692	0.866
SVM	0.883	0.8181	0.906	0.75	0.935	0.782	0.9206
Naïve Bayes	0.813	0.7	0.8484	0.5833	0.903	0.636	0.875
LR	0.883	0.8181	0.906	0.75	0.9354	0.782	0.9206
Random Forest	0.767	0.666	0.7837	0.33	0.9354	0.444	0.8529
XG Boost	0.8604	0.8	0.8787	0.66	0.9354	0.7272	0.9062

Here is a graph showing analysis for the dataset using different ML models



Based on the provided results, we can draw several conclusions:

Accuracy: SVM, Logistic Regression, and XGBoost have the highest accuracy among all classifiers, with SVM and Logistic Regression achieving 88.37% and XG Boost achieving 86.05%. Random Forest has the lowest accuracy among the classifiers, with 76.74%.

Precision: SVM, Logistic Regression, and XG Boost have similar precision for both classes (0 and 1). Decision Tree, Naive Bayes, and Random Forest show slightly lower precision compared to SVM, Logistic Regression, and XG Boost.

Recall: Random Forest has the lowest recall for Class 0, indicating it's not performing well in identifying instances of Class 0. SVM, Logistic Regression, and XG Boost have the highest recall for both classes, indicating they are better at capturing instances of both classes.

F1-*score*: SVM, Logistic Regression, and XG Boost have higher F1-scores compared to other classifiers for both classes, suggesting a good balance between precision and recall. Random Forest has a relatively lower F1-score, especially for Class 0, indicating it might struggle with overall performance compared to other classifiers.

Based on these observations, if we prioritize accuracy, precision, recall, and F1-score equally, SVM, Logistic Regression, or XG Boost might be the preferable choices for this classification task. However, the choice of the best classifier also depends on the specific requirements and constraints of the problem domain.

VII. Conclusion

Building on the foundation of a placement prediction system using machine learning models like Random Forest, LR, Decision Tree , XG Boost, SVM there are several avenues for future work to enhance the system's capabilities and address potential shortcomings.

Here are some suggestions that can be used to enhance the effective of prediction models:

1)Incorporating Additional Features: Expand the feature set to include a wider range of factors that may influence placement outcomes, such as soft skills, domain-specific certifications, industry networking, and geographic preferences. Exploring the integration of unstructured data sources, such as student resumes, cover letters, and LinkedIn profiles, using natural language processing (NLP) techniques to extract relevant information.

2)Model Interpretability and Explainability: Investigating techniques for enhancing the interpretability and explainability of the placement prediction models, particularly important for stakeholders to trust and understand the predictions. *E*xploring model-agnostic interpretability methods, such as SHAP (SHapley Additive exPlanations) values and LIME (Local Interpretable Model-agnostic Explanations), to provide insights into the decision-making process of the models.

3) and Adaptive Models: Developing dynamic and adaptive placement prediction models that can continuously learn and adapt to changing student profiles, industry trends, and job market dynamics. Implementing online learning techniques that allow the model to update its predictions in real-time as new data becomes available, ensuring its relevance and accuracy over time.

4)Personalized Recommendations and Interventions: Leveraging advanced machine learning techniques, such as collaborative filtering and reinforcement learning, to personalize placement recommendations for individual students based on their unique characteristics and career aspirations. Designing intervention strategies to provide targeted support and guidance to students who may be at risk of not securing placements, such as recommending additional skill development opportunities or networking events.

5)Evaluation and Impact Assessment: Conducting rigorous evaluation studies to assess the effectiveness and impact of the placement prediction system on student outcomes, such as placement rates, job satisfaction, and career advancement. Utilizing experimental design methodologies, such as randomized controlled trials (RCTs) or quasi-experimental designs, to measure the causal effects of the system's interventions on student success metrics.

6)Ethical and Fairness Considerations: Investigating methods for ensuring fairness and mitigating bias in the placement prediction models, particularly concerning sensitive attributes such as gender, race, and socioeconomic status. Implementing fairness-aware machine learning techniques, such as fairness constraints and adversarial debiasing, to promote equity and inclusivity in the placement process.

7)Collaboration and Integration: Foster collaboration between educational institutions, industry partners, and government agencies to enhance the predictive accuracy and relevance of the placement prediction system. Integrating the placement prediction system with existing career services platforms, learning management systems, and student advising tools to streamline the placement process and improve user experience.

By exploring these future directions, the placement prediction system can evolve into a more sophisticated and impactful tool for guiding students' career trajectories, supporting educational institutions in their mission to foster student success, and facilitating better alignment between talent supply and demand in the industry.

VIII. References

- *Classification Model of Prediction for Placement of Students*: I.J.Modern Education and Computer Science, 2013, 11, 49-56 Published Online November 2013 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijmecs.2013.11.07
- *Collaborative job prediction using naïve bayes on user skillset:* Modern Education and Computer Science, 2015, 11, 52-58 Published Online October 2015 in MECS.
- *Incorporating Features Learned by an Enhanced Deep Knowledge Tracing Model for STEM/Non-STEM Job Prediction:* Chun-Kit Yeung, Dit-Yan Yeung Published: 6 May 2019,International Artificial Intelligence in Education Society 2019
- *Model Construction Using ML for Prediction of Student Placement:* International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 10 Issue VI June 2022
- *Prediction of Final Result and Placement of Students using Classification Algorithm:* Neelam Naik, Seema Purohit, October 2012, International Journal of Computer Applications
- Sultana, Jabeen, M. Usha Rani, and M. A. H. Farquad. *"Student's performance prediction using deep learning and data mining methods."* Int. J. Recent Technol. Eng 8.1S4 (2019): 1018-1021.
- Manvitha, Pothuganti, and Neelam Swaroopa. *"Campus placement prediction using supervised machine learning techniques."* International Journal of Applied Engineering Research 14.9 (2019): 2188-2191 .
- *https://www.kaggle.com/benroshan/factors-affecting-campus-placement.*
- *https://www.kaggle.com/datasets/koshikasaiprasad/student-placement-data/code.*
- *https://www.databricks.com/glossary/machine-learning-models*
- *https://www.analyticsvidhya.com/blog/2021/06/linear-regression-in-machine-learning/*
- *https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/*
- *https://www.javatpoint.com/machine-learning-random-forest-algorithm*
- *https://www.turing.com/kb/an-introduction-to-naive-bayes-algorithm-for-beginners*